ELSEVIER
Stochastic Processes and their Applications 102 (2002) 1–23
stochastic processes and their applications
www.elsevier.com/locate/spa

Large deviations and fast simulation in the presence of boundaries

Søren Asmussen[a,*], Pascal Fuckerieder[b], Manfred Jobmann[c], Hans-Peter Schwefel[d]

[a] *Mathematical Statistics, Centre of Mathematical Sciences, Lund University, Box 118, S-221 00 Lund, Sweden*
[b] *Ernst-Haeckel-Str. 61, D-80999 München, Germany*
[c] *Institut für Informatik, Technische Universität München, D-80290 München, Germany*
[d] *Siemens AG, Hofmannstr. 51, D-81539 München, Germany*

Received 28 June 2000; received in revised form 1 March 2002; accepted 12 April 2002

Abstract

Let $\tau(x) = \inf\{t > 0: Q(t) \geqslant x\}$ be the time of first overflow of a queueing process $\{Q(t)\}$ over level x (the buffer size) and $z = \mathbb{P}(\tau(x) \leqslant T)$. Assuming that $\{Q(t)\}$ is the reflected version of a Lévy process $\{X(t)\}$ or a Markov additive process, we study a variety of algorithms for estimating z by simulation when the event $\{\tau(x) \leqslant T\}$ is rare, and analyse their performance. In particular, we exhibit an estimator using a filtered Monte Carlo argument which is logarithmically efficient whenever an efficient estimator for the probability of overflow within a busy cycle (i.e., for first passage probabilities for the unrestricted netput process) is available, thereby providing a way out of counterexamples in the literature on the scope of the large deviations approach to rare events simulation. We also add a counterexample of this type and give various theoretical results on asymptotic properties of $z = \mathbb{P}(\tau(x) \leqslant T)$, both in the reflected Lévy process setting and more generally for regenerative processes in a regime where T is so small that the exponential approximation for $\tau(x)$ is not a priori valid.

Keywords: Buffer overflow; Exponential change of measure; Filtered Monte Carlo; Importance sampling; Lévy process; Local time; Queueing theory; Rare event; Reflection; Regenerative process; Saddlepoint

* Corresponding author.
E-mail addresses: asmus@maths.lth.se (S. Asmussen), pascal@fuckerieder.de (P. Fuckerieder), jobmann@informatik.tu-muenchen.de (M. Jobmann), hans.schwefel@icn.siemens.de (H.-P. Schwefel).
URL: http://www.maths.lth.se/matstat/staff/asmus

0304-4149/02/$ - see front matter
PII: S0304-4149(02)00152-7

1. Introduction

Let $\{Q(t)\}$ be a queueing process, say the queue length, the workload or the fluid buffer content in continuous time, or the waiting time of successive customers in discrete time, and define $\tau(x) = \inf\{t > 0: Q(t) \geqslant x\}$. We are interested in evaluating

$$z = \mathbb{P}(\tau(x) \leqslant T)$$

by simulation in a situation where both x and T are large but T is small compared to x in the sense that the event $\{\tau(x) \leqslant T\}$ is rare, i.e., z is small. This is a highly relevant problem in telecommunications where one identifies x with the buffer size and $\tau(x)$ with the time of the first buffer overflow, and the values of z of interest can be as small as 10^{-9}.

For simplicity, we will assume for most of the paper that $\{Q(t)\}$ is the reflected version of a Lévy process $\{X(t)\}$ (the netput process) and that $Q(0) = X(0) = 0$, i.e.,

$$Q(t) = X(t) + L(t) \quad \text{where } L(t) = -\inf\{X(s): 0 \leqslant s \leqslant t\}. \tag{1.1}$$

Here $L(t)$ is commonly referred to as the local time, and we denote by

$$\kappa(\alpha) = \frac{1}{t} \log \mathbb{E}\, \mathrm{e}^{\alpha X(t)}$$

the Lévy exponent. See, e.g., Bertoin (1990) or Asmussen (2000, Chapter II) for some relevant background. This set-up covers simple queueing models like the M/M/1 queue length or the M/G/1 workload (and, in discrete time, where $\{X(t)\}$ is a random walk, the GI/G/1 waiting time and many imbedded Markov chains in continuous-time queues; however, we will stick to continuous time notation). We note additionally that the discussion is easily extended to the case $Q(0) > 0$ as well as to Markov-modulated models, see Section 6.

An important characteristics of the problem we study is that it deals with the queueing process itself rather than the netput process for which there is an extensive literature on rare events problems associated with $\omega(x) = \inf\{t > 0: X(t) \geqslant x\}$. Note in this connection that

$$\mathbb{P}(\tau(x) \leqslant T) \geqslant \mathbb{P}(Q(T) \geqslant x) = \mathbb{P}(\omega(x) \leqslant T) \geqslant \mathbb{P}(X(T) \geqslant x) \tag{1.2}$$

(the equality in the middle follows from the well-known fact that $Q(T)$ has the same distribution as $\max_{0 \leqslant t \leqslant T} X(t)$, cf. Asmussen (1987) III. 7–8). An important technique in the literature relating to $\mathbb{P}(\omega(x) \leqslant T)$ or $\mathbb{P}(X(T) \geqslant x)$ is exponential change of measure which amounts to considering θ with $\kappa(\theta) < \infty$ and letting $\mathbb{P}_\theta, \mathbb{E}_\theta$ refer to the case where the Lévy exponent is changed from $\kappa(\alpha)$ to $\kappa_\theta(\alpha) = \kappa(\theta + \alpha) - \kappa(\theta)$. Of particular importance in the case $\kappa'(0) < 0$ of negative drift is the case where $\theta = \gamma$, the solution > 0 of $\kappa(\gamma) = 0$. The relevant exponential change of measure can often be identified via a large deviations argument identifying the most likely path leading to the rare event. See Section 2 for more detail.

More recently, counterexamples indicating that the scope of this approach is limited have started to appear, see in particular Glasserman and Kou (1995a), Glasserman and Wang (1997) and Asmussen et al. (2000). The example of Glasserman and Kou (1995a) is of particular relevance for the present paper because the role of reflecting

boundaries is crucial. It shows (in the setting of two-dimensional random walks) that even if the most likely path can be the same for the reflected and the unrestricted process, then the efficient estimator for the unrestricted process may not work at all for the reflected one. We will present one further instant of this phenomenon when analysing Algorithm I below (the first of Algorithms I–IV discussed in this paper); this example is maybe conceptually simpler by being one dimensional. One of the main contributions of this paper is to present an algorithm (Algorithm IV) which in simple cases resolves this problem by means of a conditioning argument involving regenerative cycles. It allows in fair generality to reduce the problem of efficient simulation of large deviations probabilities for a reflected process to that of finding efficient estimators within the cycle, a problem which typically only involves the unrestricted netput process. We remark in this connection that this approach is different from a conversion from unrestricted processes to reflected ones which has been extensively used in the simulation literature (see, e.g., the survey in Asmussen and Rubinstein (1995) and references there), namely to express the tail of the stationary r.v. $Q(\infty)$ in the reflected process as a first passage probability in the unrestricted one; in the setting of (1.2), the identity is $\mathbb{P}(Q(\infty) \geqslant x) = \mathbb{P}(\omega(x) < \infty)$ and is obtained by letting $T \to \infty$.

In Section 2, we give a somewhat more elaborate discussion of rare events behaviour. Also some limit results of independent interest for rare events in regenerative processes are given. In the setting of reflected Lévy processes, the discussion of Section 2 leads to the suggestion of Algorithms I–III for simulating z efficiently in Section 3. However, numerical illustrations and theoretical results show that these algorithms do not meet the optimality concept of logarithmic efficiency discussed in the rare events simulation literature. The positive results are then in Section 4, where we describe Algorithm IV (the one involving conditioning via regenerative cycles as mentioned above) and show certain optimality properties. We point out also that this algorithm has potential beyond the reflected Lévy process setting (but see the discussion of the difficulties arising in cases such as the problem studied by Glasserman and Kou (1995a)). Finally, some of the more technical proofs are deferred to Section 5 (a key tool is sample path large deviations), and some concluding remarks and extensions are in Section 6.

2. Preliminaries

We start by three examples.

Example 2.1. The M/M/1 queue length process corresponds to $X(t) = N_\lambda(t) - N_\mu(t)$ where N_λ, N_μ are independent Poisson processes with intensities λ, resp. μ (λ denotes the arrival intensity and μ the service intensity). Further

$$\kappa(\alpha) = \log \mathbb{E}\, \mathrm{e}^{\alpha X(1)} = \lambda(\mathrm{e}^{\alpha} - 1) + \mu(\mathrm{e}^{-\alpha} - 1)$$

so that

$$\begin{aligned}\kappa_\theta(\alpha) &= \kappa(\theta + \alpha) - \kappa(\theta) = \lambda(\mathrm{e}^{\alpha+\theta} - \mathrm{e}^{\theta}) + \mu(\mathrm{e}^{-\alpha-\theta} - \mathrm{e}^{-\theta}) \\ &= \lambda_\theta(\mathrm{e}^{\alpha} - 1) + \mu_\theta(\mathrm{e}^{-\alpha} - 1),\end{aligned}$$

where $\lambda_\theta=\lambda e^{\theta}$, $\mu_\theta=\mu e^{-\theta}$. Thus $\mathbb{P}_\theta$ corresponds to a M/M/1 queue with arrival intensity λ_θ and service intensity μ_θ. It follows also that $\gamma=-\log\rho$ where $\rho=\lambda/\mu$, and hence $\lambda_\gamma=\mu$, $\mu_\gamma=\lambda$. The quantity $\kappa'(\gamma)$ plays an important role in the following, and one gets easily $\kappa'(\gamma)=\mu-\lambda$.

The busy cycle can be defined as either of

$$C'=\inf\{t>0\colon Q(t)=1,\ Q(t-)=0\,|\,Q(0)=1\},$$

$$C''=\inf\{t>0\colon Q(t)=0,\ Q(t-)=1\,|\,Q(0)=0\}.$$

Example 2.2. The M/G/1 workload process corresponds to $X(t)=\sum_{i=1}^{N_\lambda(t)}U_i-t$ where $U_1,U_2,\ldots$ are i.i.d. service times independent of N_λ. Further $\kappa(\alpha)=\lambda(\mathbb{E}\,e^{\alpha U}-1)-t$ and it follows easily along the lines of Example 2.1 that $\mathbb{P}_\theta$ corresponds to a M/G/1 queue with arrival intensity $\lambda_\theta=\lambda\mathbb{E}\,e^{\theta U}$ and service time distribution obtained by exponential tilting with θ,

$$\mathbb{P}_\theta(U\in du)=\frac{e^{\theta u}}{\mathbb{E}\,e^{\theta U}}\mathbb{P}(U\in du).$$

In the M/M/1 case, this means that $\mathbb{P}_\theta$ corresponds to a M/M/1 queue with arrival intensity $\lambda_\theta=\lambda+\theta$ and service intensity $\mu_\theta=\mu-\theta$, i.e., an additive change of intensities rather than a multiplicative one as for the queue length. However, λ_γ and μ_γ are the same for the two cases (for the M/M/1 workload, $\gamma=\mu-\lambda$ and $\kappa'(\gamma)=1/\rho-1$; in the general M/G/1 case, γ and $\kappa'(\gamma)$ have to be computed numerically).

As in the preceding example, a busy cycle is composed of a busy period and an idle period, the order of which is unimportant.

Example 2.3. If $\{X(t)\}$ is Brownian motion with drift $-\mu$ and variance constant σ^2, then $\{Q(t)\}$ is reflected Brownian motion with the same parameters. This process occurs widely as an approximation to in part substantially more complicated queueing processes (see, e.g., Whitt, 2002). One has $\kappa(\alpha)=-\alpha\mu+\alpha^2\sigma^2/2$ and it follows easily along the lines of Example 2.1 that $\mathbb{P}_\theta$ corresponds to changing the drift to $\mu_\theta=\mu+\theta/\sigma^2$, whereas the variance remains unaffected. Further, $\gamma=2\mu$, $\kappa'(\gamma)=\mu$.

The choice of a regenerative cycle is less canonical than in the two preceding examples, but one can take, e.g.,

$$C=\inf\{t>0\colon Q(t)=0,\ \tau(1)<t\,|\,Q(0)=0\}$$

("up to 1 from 0 and back to 0 again").

The following result, covering all three examples, will be used in the following:

Lemma 2.4. *For any Lévy process with $\kappa'(0)<0$ and $\kappa'(\gamma)<\infty$, there exists a constant K such that for any choice of the regenerative cycle C for $\{Q(t)\}$,*

$$\mathbb{P}(\tau(x)<C)\sim K\mathbb{E}C\,e^{-\gamma x},\quad x\to\infty.$$

Proof. The asymptotic exponentiality follows by exponential change of measure, cf. Asmussen (1999), and the independence of K of C is shown in Lemma 2.6(ii) of Asmussen (1998). □

The following results give the order of magnitude of the rare event probability z. The proofs are more technical and deferred to Section 5.

Theorem 2.5. *Let $\{Q(t)\}$ be any regenerative process with generic cycle C, let $\bar{f}(x)=\mathbb{P}(\tau(x)<C)$, $\bar{f}(t;x)=\mathbb{P}(\tau(x)\leqslant t<C)$ and assume that $T(x)\uparrow\infty$, $x\to\infty$, in such a way that*

$$\lim_{x\to\infty}\bar{f}(x)T(x)=0,\quad \lim_{x\to\infty}\frac{\bar{f}(\varepsilon T(x);x)}{\bar{f}(x)}=1 \tag{2.1}$$

for all $\varepsilon>0$. Then $\mathbb{P}(\tau(x)\leqslant T(x))\sim\bar{f}(x)T(x)/\mathbb{E}C$.

The result should be compared with the standard exponential approximation for $\tau(x)$, stating that $\mathbb{P}(\tau(x)\leqslant T(x))\to 1-\mathrm{e}^{-y}$ provided $\bar{f}(x)T(x)/\mathbb{E}C\to y$, see Keilson (1966), Gnedenko and Kovalenko (1989) and Glasserman and Kou (1995b) [basically, Theorem 2.5 is a more informative version of this statement for the case $y=0$].

Corollary 2.6. *Under the conditions of Lemma* 2.4, $\mathbb{P}(\tau(x)\leqslant T(x))\sim\mathrm{e}^{-\gamma x}T(x)K$ *provided* $\mathrm{e}^{-\gamma x}T(x)\to 0$ *and* $T(x)/x\to\infty$.

Let

$$\kappa^*(m)=\sup_{\theta}\left[\theta m-\kappa(\theta)\right]$$

be the large deviations rate function (see, e.g., Dembo and Zeitouni (1998)). When needed (as to avoid technicalities in the following theorem), we will tacitly assume that $\kappa(\cdot)$ is steep in the sense of Dembo and Zeitouni (1998, p. 44) which implies that the sup is attained for the unique $\theta=\theta(m)$ satisfying the saddlepoint equation $\kappa'(\theta)=m$. For example, in the Brownian case (Example 2.3) with $\sigma^2=1$, one has $\kappa^*(x)=(x+\mu)^2/2$, whereas for the M/M/1 queue length process (Example 2.1) one gets

$$\kappa^*(x)=x\log g(x)-\lambda(g(x)-1)-\mu(g(x)^{-1}-1)$$

$$\text{where } g(x)=\frac{1}{2\lambda}(x+\sqrt{x^2+4\lambda\mu}).$$

Theorem 2.7. *Consider a reflected Lévy process $\{X(t)\}$, and assume $x/T(x)\to m$ where $m>\kappa'(0)>0$ or $\kappa'(0)<0$, $m>\kappa'(\gamma)$. Then*

$$\frac{1}{x}\log\mathbb{P}(\tau(x)\leqslant T(x))\to-\kappa^*(m)/m.$$

Note that it is well known (Anantharam, 1988) that the same conclusion holds for the (smaller, cf. (1.2)) probabilities $\mathbb{P}(Q(T(x))\geqslant x)$ and $\mathbb{P}(X(T(x))\geqslant x)$. Thus, basically

$\{\tau(x) \leqslant T(x)\}$ occurs by $\{X(t)\}$ reaching level x at time $T(x)$ and not before and, as the proof will show, level 0 is left almost instantaneously at $t = 0$ so that $Q(t)$ and $X(t)$ are almost identical when $t \leqslant T(x)$.

Now return to the problem of evaluating $z = z(x) = \mathbb{P}(\tau(x) \leqslant T)$ by simulation. The standard Monte Carlo procedure is to determine a r.v. $Z = Z(x)$ which has mean z and can be generated by simulation. Then N i.i.d. replicates $Z_1, \ldots, Z_N$ of Z are generated and the point estimator is the empirical mean

$$\bar{Z}_N = \frac{1}{N} \sum_{i=1}^{N} Z_i$$

with associated 95% confidence interval $\bar{Z}_N \pm 1.96\hat{\sigma}/N^{1/2}$ where $\hat{\sigma}^2$ is the empirical variance of $Z_1, \ldots, Z_N$.

Since we assume z to be small, the problem is one of rare events simulation, see e.g., the survey papers Heidelberger (1995) and Asmussen and Rubinstein (1995) where one ideally looks for estimators $Z(x)$ which have bounded relative error,

$$\limsup_{x \to \infty} \frac{\mathbf{Var}(Z(x))}{z(x)^2} < \infty$$

or at least has the slightly weaker property

$$\limsup_{x \to \infty} \frac{\mathbf{Var}(Z(x))}{z(x)^{2-\varepsilon}} < \infty \quad \text{for all } \varepsilon > 0$$

of logarithmic efficiency discussed in Heidelberger (1995) and Asmussen and Rubinstein (1995). In many examples, this can be achieved by importance sampling where one uses a change of measure making the rare event $A(x)$ (say) more likely. That is, if $\tilde{\mathbb{P}}$ is the changed measure ($\tilde{\mathbb{P}}$ could depend on x), then $Z(x) = \mathrm{d}\mathbb{P}/\mathrm{d}\tilde{\mathbb{P}} \cdot I(A(x))$. More specifically, it is well known that it often works to take $\tilde{\mathbb{P}}$ close in an asymptotic sense to the conditional distribution $\mathbb{P}(\cdot \mid A(x))$ given the rare event as possible (but see the Introduction for counterexamples!). Via conditioned limit theorems such as in Asmussen (1982), this leads into exponential change of measure as surveyed above. The likelihood ratio is then

$$\left.\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}_\theta}\right|_T = \exp\{-\theta X(T) + T\kappa(\theta)\}$$

(T may be constant or a stopping time like $\tau(x)$ or $\tau(x) \wedge T$).

Three main examples of relevance for the following are:

1. $A(x) = \{X(T) \geqslant x\}$ where $x > T\kappa'(0)$. Then the saddlepoint method suggests to choose θ such that $\mathbb{E}_\theta X(T) = x$, i.e., θ is the solution of $T\kappa'(\theta) = x$. The resulting algorithm is logarithmically efficient as $x \to \infty$ if $T = T(x)$ varies with x in such a way that x/T has a limit $m > \kappa'(0)$, cf. Bucklew et al. (1990).
2. $A(x) = \{\omega(x) \leqslant T\}$ where $x > T\kappa'(0) > 0$ or $\kappa'(0) < 0$, $x > T\kappa'(\gamma)$. Again, the saddlepoint method suggests to choose θ such that $\mathbb{E}_\theta X(T) = x$, i.e. $T\kappa'(\theta) = x$. The

resulting algorithm is logarithmically efficient as $x \to \infty$ if $T = T(x)$ varies with x in such a way that x/T has a limit m, cf. Asmussen (2000) X.4.

3. $A(x) = \{\omega(x) < \infty\}$ (assuming $\kappa'(0) < 0$ to make $A(x)$ rare). Here the choice $\theta = \gamma$ gives relative bounded error, cf., e.g., Asmussen and Rubinstein (1995).

As a fourth example, essentially a small variant of Item 3, consider $A(x) = \{\tau(x) < C\}$ where C is a regenerative cycle for $\{Q(t)\}$. For a simple reflected random walk with the starts of cycles being defined as the return times to 0, exponential choice of measure with $\theta = \gamma$ gives bounded relative error (see Asmussen and Rubinstein, 1995, p. 436). With some care, this holds quite generally, but that care is indeed needed, is illustrated via the following example which also may serve as an introductory warning that the presence of boundaries may cause trouble:

Example 2.8. Consider as in Example 2.1 the M/M/1 queue length, and recall the two choices C', C'' of the busy cycles. If we choose C', corresponding to starting the busy cycle with one customer who just arrived, the paths of $\{Q(t)\}$ and $\{1+X(t)\}$ coincide up to $t = C'$ when $\tau(x) < C'$, and the estimator corresponding to importance sampling governed by $\mathbb{P}_\gamma$ is

$$Z'(x) = \rho^{x-1} I(\tau(x) < C') = \rho^{X(\tau(x))} I(\tau(x) < C') = \rho^{x-1} I(X(\omega^*(x)) = x - 1),$$

where $\omega^*(x) = \inf\{t > 0 : X(t) \in \{-1, x-1\}\}$. From this it is immediate that the second moment is of order ρ^{2x} and hence (since it is well known that $\mathbb{P}(\tau(x) < C)$ is of order ρ^x) that the estimator has bounded relative error.

Now apply instead the choice C'', corresponding to starting the busy cycle with the system empty. Using again importance sampling from $\mathbb{P}_\gamma$, the estimator is

$$Z''(x) = \rho^{X(\tau(x))} I(\tau(x) < C'') = \rho^x \rho^{-L(\tau(x))} I(\tau(x) < C''),$$

where $L(\tau(x))$ coincides with the number M of fictitious service events during the idle period on $\{\tau(x) < C''\}$. Conditionally upon an idle period of length x, M is Poisson(λx) w.r.t. $\mathbb{P}_\gamma$, and hence

$$\mathbb{P}_\gamma(M = m) = \int_0^\infty \mu \, \mathrm{e}^{-\mu x} \, \mathrm{e}^{-\lambda x} \, \frac{(\lambda x)^m}{m!} \, \mathrm{d}x = \frac{\mu \lambda^m}{(\mu + \lambda)^{m+1}} = \frac{\rho^m}{(1+\rho)^{m+1}}.$$

Therefore $\mathbb{E}_\gamma \rho^{-2M} < \infty$ if and only if $\rho(1+\rho) > 1$, i.e. $\rho > 0.62$, and since M and $\{\tau(x) < C''\}$ are independent, this is also the condition for $Z''(x)$ to have a finite second moment (then indeed one has bounded relative error).

An obvious modification applies to produce an estimator which has bounded relative error for all $\rho < 1$: do not use importance sampling in the idle period but turn it on at the arrival time of the first customer. Similarly, in the Brownian case in Example 2.3 one should not use importance sampling in the whole cycle but only after (say) level 1 has been hit (we omit the details of calculation supporting this statement).

3. Algorithms–first attempts

As discussed above, importance sampling aims at making the rare event in some sense 'typical' in the changed distribution. We here suggest three ways to do this and in the next section a fourth algorithm which has a somewhat different structure.

The first idea is to impose a linear drift to take $\{Q(t)\}$ to level x at time T (the motivation for the procedure leading to $Z_{\rm I}'$ may not be clear at this stage, but is motivated from Anantharam (1988) and further discussed in Section 6).

Algorithm I. (i) If $\kappa'(0) > 0,\ T\kappa'(0) < x$ or $\kappa'(0) < 0,\ T\kappa'(\gamma) < x$, choose θ such that $T\kappa'(\theta) = x$. Simulate from $\mathbb{P}_\theta$ until $T \wedge \tau(x)$ and use the estimator

$$Z_{\rm I} = Z_{\rm I}(x) = \exp\{-\theta X(\tau(x)) + \tau(x)\kappa(\theta)\} I(\tau(x) \leqslant T).$$

(ii) If $\kappa'(0) < 0,\ T\kappa'(\gamma) > x$, simulate without importance sampling until $t(x) = T - x/\kappa'(\gamma)$ (or just $\tau(x)$ if $\tau(x) < t(x)$). If $\tau(x) > t(x)$, simulate from $\mathbb{P}_\gamma$ in $(t(x), T \wedge \tau(x)]$. The overall estimator is

$$Z_{\rm I}' = Z_{\rm I}'(x) = I(\tau(x) \leqslant t(x)) + \exp\{-\gamma[X(\tau(x)) - X(t(x))]\} I(t(x) < \tau(x) \leqslant T).$$

The second idea is to make T central in the distribution of $\tau(x)$ as follows:

Algorithm II. Choose θ such that $\mathbb{E}_\theta \tau(x) = T$. Simulate from $\mathbb{P}_\theta$ until $T \wedge \tau(x)$ and use the estimator

$$Z_{\rm II} = Z_{\rm II}(x) = \exp\{-\theta X(\tau(x)) + \tau(x)\kappa(\theta)\} I(\tau(x) \leqslant T).$$

This choice of θ is definitely more intricate than in Algorithm I, but feasible in reasonable generality. In fact, Asmussen and Kella (2001) give a general formula for the expected value of a general stopping time for a reflected Lévy process. E.g., for the queue length process in M/M/1, this leads to

$$\mathbb{E}\tau(x) = \frac{(\rho^{-x}\,\mathrm{e}^{-2x\gamma} - 1) - x(1 - \rho\,\mathrm{e}^{2\gamma})}{\mu(1 - \rho\,\mathrm{e}^{2\gamma})^2}.$$

The approach generalizes to Markov-modulated systems, see Asmussen et al. (2002) and references there.

The performance of Algorithms I and II is illustrated in Table 1 for the M/M/1 queue length process (Example 2.1) with $\lambda = 0.5$, $\mu = 1$ and (as for Fig. 1 and

Table 1
Estimators for different M/M/1 queues using Algorithms I and II

x	T	z	$\hat{z}_{\rm I}$	$\mathrm{Var}(\hat{z}_{\rm I})/\hat{z}_{\rm I}^2$	$\hat{z}_{\rm II}$	$\mathrm{Var}(\hat{z}_{\rm II})/\hat{z}_{\rm II}^2$
10	10	3.0×10^{-4}	3.9×10^{-4}	2.1	4.2×10^{-4}	6.7
15	15	5.7×10^{-6}	6.6×10^{-6}	2.9	6.8×10^{-6}	6.5
20	20	1.1×10^{-7}	1.3×10^{-7}	14.9	1.2×10^{-7}	4.3
25	25	2.2×10^{-9}	2.4×10^{-9}	6.2	2.5×10^{-9}	4.6
30	30	4.3×10^{-11}	5.1×10^{-11}	15.7	6.4×10^{-11}	12.7

Table 2 below) 1000 replications. As comparison, we included also exact values of the rare event probability $z(x)$ which were obtained from the algorithm of Asmussen et al. (2002) (the method behind is to obtain the Laplace transform of $\tau(x)$ explicitly by optional stopping of a martingale of Kella and Whitt (1992), and to invert the Laplace transform numerically).

It is seen that the differences between Algorithms I and II are minor (the values of θ were also almost identical), and both appear to produce reliable results. However, the fluctuations in the estimated relative error appear surprisingly large, and in fact the following theoretical results give an explanation of this and shows that one should not trust Algorithm I at least for stable queues (in view of the similarities between Algorithms I and II, we have omitted a theoretical analysis of Algorithm II). For the proofs, see Section 5.

Theorem 3.1. *Consider the limit* $x/T(x) \to m > \kappa'(0)$. *Then*:

(a) *Algorithm* I(i) *is logarithmic efficient if* $\kappa(-\theta) < \kappa(\theta)$;
(b) *Algorithm* I(i) *is not logarithmic efficient if* $\kappa(-\theta) > \kappa(\theta)$.

On the positive side:

Corollary 3.2. *If* $\kappa'(0) > 0$, *then there exists* $m_1 > \kappa'(0)$ *such that for all* $m \in (\kappa'(0), m_1)$, *Algorithm* I(i) *is logarithmic efficient in the limit* $x/T(x) \to m$.

However, in the stable case the result is disappointing:

Corollary 3.3. *If* $\kappa'(0) < 0$, *then there exists* $m_2 > \kappa'(\gamma)$ *such that for no* $m \in (\kappa'(\gamma), m_2)$, *Algorithm* I(i) *is logarithmic efficient in the limit* $x/T(x) \to m$. *In the case of reflected Brownian motion with drift or the* $M/M/1$ *queue length process*, $m_2 = \infty$.

We show below (Example 4.3) that nevertheless an efficient estimator based upon the same exponential choice of measure can be produced.

The third algorithm is more sophisticated and uses a regenerative argument, viewing the exceedance of level x as result of independent trials in the busy cycles. This is motivated by the crucial role this point of view plays in extreme value theory, cf. Asmussen (1999) and references there. In fact, it appears at a first sight that this algorithm is the one which is closest to the idea of simulating from a distribution close to $\mathbb{P}(\cdot \mid \tau(x) \leqslant T)$. Let $C_1, C_2, \ldots$ be the consecutive busy cycles and

$$M(T) = \inf\{n\colon C_1 + \cdots + C_n > T\}$$

(here by convention $C_1 + \cdots + C_n = 0$ when $n = 0$). Note that $M(T)$ is of order $T/\mathbb{E}C$. The importance sampling distribution $\tilde{\mathbb{P}}$ is determined as follows. For each cycle, a coin is flipped coming up heads with probability p where $p = \mathbb{E}C/T$ (i.e., the expected number of heads is approximatively 1). If tails come up, the cycle is simulated without

change of measure, if heads come up $\mathbb{P}_\gamma$ is used until the queueing process reaches level x or the cycle is completed (whatever happens first). We get:

Algorithm III. At the start of cycle i, generate a 0–1 variable U_i with $\mathbb{P}(U_i=1)=p$. If $U_i=0$, simulate cycle i without importance sampling. If $U_i=1$, simulate from $\mathbb{P}_\gamma$ until $\{Q(t)\}$ reaches level x or the cycle is completed, and let L_i be the likelihood at that time. Define further $V_i=1$ if the queue length reaches level x in the cycle, $V_i=0$ otherwise. The estimator is

$$Z_{\mathrm{III}}=Z_{\mathrm{III}}(x)=I(\tau(x)\leqslant T)\prod_{i=1}^{M(T)\wedge K}\frac{1}{(1-p)^{1-U_i}}\left[\frac{1}{p}L_i\right]^{U_i},$$

where $K=\inf\{i\colon V_i=1\}$.

The empirical results on the performance of this estimator (omitted here; see Frantz, 2000) were, however, disappointing and in fact, the following example shows that the relative error can be huge so that Algorithm III should not be used to estimate the probability of a buffer overflow.

Example 3.4. Take $x=21$, $T=10000$, $\lambda=0.5$, $\mu=1.0$ so that $\mathbb{P}(U_i=1)=4\times10^{-4}$. In one of our numerical experiments, an overflow occurred in cycle C_{1000} without importance sampling ($U_{1000}=0$), so that the run is ended in cycle C_{1000}. In the previous 999 cycles, importance sampling was turned on once, but no overflow occurred. This sample path gives a likelihood ratio of

$$\frac{1}{(1-4\times10^{-4})^{999}p}\,\mathrm{e}^{-\log(0.5/1.0)}\approx1864 \tag{3.1}$$

and therefore has a large contribution to the estimator. We will show that also the contribution to the second moment is huge. If $\gamma(T)$ is the probability of an overflow in the period $[0,T]$, then by Theorem 2.5 $\gamma(T/2)\sim\gamma(T)/2$. Hence we get asymptotically that

$$\begin{aligned}\mathbb{E}Z_{\mathrm{III}}^2&\geqslant\left(\frac{Tp}{2\mathbb{E}C}(1-\rho)(1-p)^{T/2\mathbb{E}C-1}\right)\gamma(T/2)\left(\frac{\rho}{(1-p)^{T/\mathbb{E}C-1}p}\right)^2\\&\geqslant\frac{T}{2p\mathbb{E}C}(1-\rho)\rho^2\gamma(T/2)\approx\frac{T^2}{2\mathbb{E}C^2}(1-\rho)\rho^2\frac{1}{2}\gamma(T)\\&=T^2\gamma(T)\frac{(1-\rho)\rho^2}{4\mathbb{E}C^2}.\end{aligned}$$

In fact, note that (a) $T/2\mathbb{E}C$ is the average number of cycles in a time interval $[0,T/2]$, (b) for large buffers no overflow in a cycle without importance sampling occurs with probability $(1-\rho)$, (c) $p(1-p)^{T/2\mathbb{E}C-1}$ corresponds to exactly one cycle being simulated using importance sampling. Considering the contribution from the specific sample path, we therefore get the lower bound

$$\frac{\sqrt{T^2\gamma(T)((1-\rho)\rho)/4\mathbb{E}C^2-\gamma(T)^2}}{\gamma(T)}\approx12826$$

for the relative error.

4. An efficient algorithm

The final algorithm takes advantage of the fact that an efficient algorithm for the simulation of $\mathbb{P}(\tau(x) \leqslant C)$ is available, cf. Item 3 at the end of Section 2. This is then combined with a conditioning argument using the renewal representation and thus the algorithm contains the ingredient of conditional Monte Carlo, more precisely extended conditional Monte Carlo in the sense of Bratley et al. (1987) (see also Glasserman, 1996).

Algorithm IV. We use the representation

$$\mathbb{P}(\tau(x) \leqslant T) = \int_0^T \bar{f}(T-t;x)U(\mathrm{d}t;x), \tag{4.2}$$

where $\bar{f}(t;x) = \mathbb{P}(\tau(x) \leqslant t < C)$ as in Theorem 2.5 and

$$U(A;x) = \sum_{k=0}^{\infty} \mathbb{P}(C_1 + \cdots + C_k \in A,\ \tau(x) > C_1 + \cdots + C_k)$$

(to obtain (4.2), condition upon the time $t = C_1 + \cdots + C_k$ where the cycle containing $\tau(x)$ starts).

Simulate first a cycle $\{Q(t)\}_{0 \leqslant t \leqslant C}$ with importance sampling from $\mathbb{P}_\gamma$ implemented such that $L^*(x)I(\tau(x) < C)$ is an estimator of $\mathbb{P}(\tau(x) < C)$ with bounded relative error (cf. Example 2.8) where $\tau^*(x) = \tau(x)$ if $\tau(x) < C$, $\tau^*(x) = \infty$ otherwise and $L^*(x)$ is the likelihood ratio at time $\tau^*(x)$. Then $\hat{\bar{f}}(t;x) = L^*(x)I(\tau^*(x) \leqslant t)$ is an unbiased estimator of $f(t;x)$. Simulate next cycles

$$\{Q(t)\}_{0 \leqslant t \leqslant C_1},\ \{Q(t)\}_{0 \leqslant t \leqslant C_2}, \ldots, \{Q(t)\}_{0 \leqslant t \leqslant C_N} \tag{4.3}$$

without importance sampling, where

$$N = \inf\left\{k : C_1 + \cdots + C_k > T \text{ or } \max_{0 \leqslant t \leqslant C_k} Q(t) \geqslant x\right\}.$$

Then for $A \subseteq [0,T]$,

$$\hat{U}(A;x) = \sum_{k=0}^{N-1} I(C_1 + \cdots + C_k \in A)$$

is an unbiased estimator of $U(A;x)$ and we can let

$$\begin{aligned} Z_{\mathrm{IV}} = Z_{\mathrm{IV}}(x) &= \int_0^T \hat{\bar{f}}(T-t;x)\hat{U}(\mathrm{d}t;x) \\ &= L^*(x) \sum_{k=0}^{N-1} I(\tau^*(x) \leqslant T - C_1 - \cdots - C_k). \end{aligned}$$

Note that if $\tau^*(x) = \infty$, then $Z_{\mathrm{IV}} = 0$ and it is not necessary to simulate (4.3) (also if $\tau^*(x) < \infty$, it may happen that $Z_{\mathrm{IV}} = 0$, namely if $\tau^*(x) > T$).

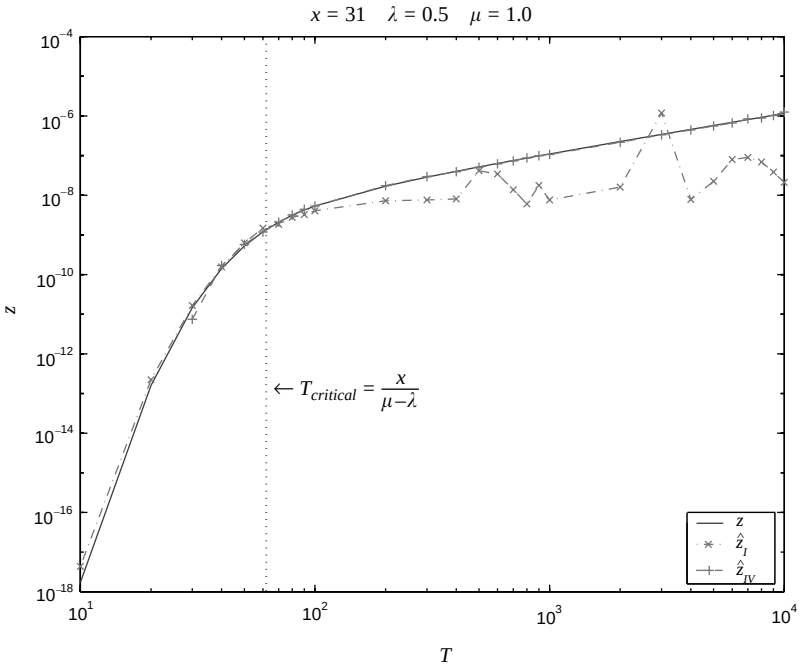

Fig. 1. Estimated probability of buffer overflow in an M/M/1 queuing system with buffer size $x = 31$, arrival rate $\lambda = 0.5$ and service rate $\mu = 1.0$. The figure shows the estimates produced by Algorithms I and IV and the exact values of z computed by transform inversion for different simulation times T.

Table 2
Estimators for different M/M/1 queues using Algorithm IV

x	T	z	$\hat{z}_{\mathrm{IV}}$	$\mathrm{Var}(\hat{z}_{\mathrm{IV}})/\hat{z}_{\mathrm{IV}}^2$
12	35	1.2×10^{-3}	1.2×10^{-3}	1.6
16	53	1.1×10^{-4}	1.2×10^{-4}	1.3
20	80	1.1×10^{-5}	1.1×10^{-5}	1.5
24	121	1.2×10^{-6}	1.2×10^{-6}	1.2
28	184	1.3×10^{-7}	1.3×10^{-7}	1.0
32	279	1.3×10^{-8}	1.4×10^{-8}	0.9
36	422	1.3×10^{-9}	1.2×10^{-9}	1.2
40	640	1.3×10^{-10}	1.3×10^{-10}	1.0
44	970	1.3×10^{-11}	1.2×10^{-11}	1.2

The performance of Algorithm IV is illustrated in Table 2 for the same M/M/1 example as in Table 1. A further numerical study is in Fig. 1, including also numbers produced by Algorithm I. The good results are confirmed by:

Theorem 4.1. *Assume* $\kappa'(0) < 0$, $\kappa'(\gamma) < \infty$, $\mathrm{e}^{-\gamma x} T(x) \to 0$ *and* $T(x)/x \to \infty$. *Then Algorithm* IV *has bounded relative error.*

Proof. According to the estimate of the rare event probability given by Theorem 2.5 and Corollary 2.6, we must show that

$$\limsup_{x\to\infty} \frac{\mathbb{E}Z_{\mathrm{IV}}^2}{(\bar{f}(x)T(x))^2} < \infty.$$

But clearly

$$\mathbb{E}Z_{\mathrm{IV}}^2 \leqslant \mathbb{E}_\gamma[L^*(x)^2;\tau^*(x)<\infty]\cdot\mathbb{E}N^2 = \mathbb{E}_\gamma[L^*(x)^2;\tau^*(x)<C]\cdot\mathbb{E}N^2.$$

By assumption, the first factor on the r.h.s. is of order $\bar{f}(x)^2$, whereas we can bound the second by the second moment of the number of renewals (starts of busy cycles) before $T(x)$ which is of order $T(x)^2$. From this the result follows. □

It is clear that the ideas behind Algorithm IV apply in considerably more general situations than the one in Theorem 4.1. To this end, assume that a family $\hat{\bar{f}}(t;x)$ of unbiased estimators of the $\bar{f}(t;x)$ is available (not necessarily obtained by importance sampling from $\mathbb{P}_\gamma$) and given such a family, define Algorithm IV* just as above. Precisely the same argument as in the proof of Theorem 4.1 yields:

Corollary 4.2. *Assume that*

$$0 \leqslant \hat{\bar{f}}(t;x) \leqslant Y(x),\ 0 \leqslant t \leqslant T$$

for some r.v.'s $Y(x)$ *satisfying*

$$\limsup_{x\to\infty} \frac{T(x)^2\mathbb{E}Y(x)^2}{\mathbb{P}(\tau(x)\leqslant T(x))^{2-\varepsilon}} < \infty$$

for all $\varepsilon > 0$ *or, equivalently,*

$$\limsup_{x\to\infty} \frac{2\log T(x) + \log[\mathbb{E}Y(x)^2]}{2\log\mathbb{P}(\tau(x)\leqslant T(x))} \geqslant 2. \tag{4.4}$$

Then Algorithm IV* *is logarithmically efficient.*

Example 4.3. In the setting of Algorithm I(i), choose again θ such that $T\kappa'(\theta) = x$, simulate $\{X(t)\}$ from $\mathbb{P}_\theta$ until $\omega(x)$ and let

$$\hat{\bar{f}}(t;x) = \exp\{-\theta X(\omega(x)) + \omega(x)\kappa(\theta)\}I(\omega(x)\leqslant t).$$

Noting that the assumptions in case (i) of Algorithm I imply $\kappa(\theta) > 0$, we can use

$$Y(x) = \exp\{-\theta x + T\kappa(\theta)\} = \exp\{-x\kappa^*(m)/m\}$$

as upper bound (recall that $x = T(x)m$); condition (4.4) then immediately follows from Theorem 2.7.

Example 4.4. We consider here a discrete two-dimensional random walk example identical to the tandem queue setting in Glasserman and Kou (1995a). The unrestricted process $\{X_n = (X_n^{(1)}, X_n^{(2)})\}_{n=0,1,2,\ldots}$ has state space $\{0, \pm 1, \pm 2, \ldots\}^2$ and the restricted one $\{Q_n = (Q_n^{(1)}, Q_n^{(2)})\}_{n=0,1,2,\ldots}$ state space $E = \{0, 1, 2, \ldots\}^2$. The transition probabilities for $\{X_n\}$ are

$$p_X((n_1, n_2), (n_1 + m_1, n_2 + m_2)) = \begin{cases} \lambda/(\lambda + \mu_1 + \mu_2) & m_1 = 1,\ m_2 = 0, \\ \mu_1/(\lambda + \mu_1 + \mu_2) & m_1 = -1,\ m_2 = 1, \\ \mu_2/(\lambda + \mu_1 + \mu_2) & m_1 = 0,\ m_2 = -1, \\ 0 & \text{otherwise.} \end{cases}$$

For $\{Q_n\}$, $p_Q((n_1, n_2), (n_1 + m_1, n_2 + m_2)) = p_X((n_1, n_2), (n_1 + m_1, n_2 + m_2))$ when $n_1 > 0, n_2 > 0$; for $n_1 = 0$ or $n_2 = 0$ certain modifications apply that are not needed in detail for the following discussion. We assume $Q_0 = X_0 = (0, 0)$, and define

$$C = \inf\{n = 1, 2, \ldots : Q_n = (0, 0) \mid Q_0 = (0, 0)\},$$

$$\tau(x) = \inf\{n = 1, 2, \ldots : Q_n^{(1)} + Q_n^{(2)} = x\},$$

$$\omega_1(x) = \inf\{n = 1, 2, \ldots : X_n^{(1)} + X_n^{(2)} = x\}.$$

We assume stability, $\lambda\mu_1 < 1,\ \lambda\mu_2 < 1$ (then $\mathbb{E}C < \infty$), and also that $\mu_2 < \mu_1$. The rare event probability under study is $z = \mathbb{P}(\tau(x) < C)$. The exponential change of measure in Glasserman and Kou (1995a) corresponds to interchanging λ and μ_2 (write $L(n)$ for the corresponding likelihood ratio at time n). As discussed in Glasserman and Kou (1995a), this change of measure is the one suggested by the large deviations approach and is in fact efficient for the unrestricted process $\{X_n\}$ but it leads to an estimator for z in the reflected process $\{Q_n\}$ which has infinite variance.

Tempting to adapt Algorithm IV*, let

$$\sigma_0 = \inf\{n > 0\colon Q_n \in E \setminus \Delta\},$$

$$\sigma_k = \inf\{n > \sigma_{k-1}\colon Q_{n-1} \in \Delta, Q_n \in E \setminus \Delta\},$$

$$N = \inf\{k\colon \tau(x) \leqslant \sigma_k \text{ or } C \leqslant \sigma_k\}, \quad x_k = x - Q_{\sigma_k}^{(1)} - Q_{\sigma_k}^{(2)},$$

where $\Delta = \{(n_1, n_2)\colon n_1 = 0 \text{ or } n_2 = 0\}$ (note that $x_k \in \{x - 1, x\}$). Simulate first the segment $X_0^*, \ldots, X_{\omega_1(x)}^*$ of an independent copy $\{X_n^*\}$ of $\{X_n\}$ with the change of measure and next the segment $X_0, \ldots, X_{\sigma_N}$ of $\{X_n\}$ without, and define

$$\delta^*(m_1, m_2) = \inf\{n\colon (m_1, m_2) + X_n^* \in \Delta\},$$

$$\hat{z} = \sum_{k=0}^{N-1} L(\omega_1(x_k)) I(\omega_1(x_k) \leqslant \delta^*(Q_{\sigma_k}^{(1)}, Q_{\sigma_k}^{(2)})).$$

Copying the calculations above now show that σ_k with $Q_{\sigma_k}^{(1)} + Q_{\sigma_k}^{(2)}$ contribute only $O(z^2)$ to the variance (this uses efficiency properties shown in Glasserman and Kou (1995a)), whereas it is less clear how to control the (rare) σ_k with either $Q_{\sigma_k}^{(1)}$ or

$Q^{(2)}_{\sigma_k}$ close to x, and how Algorithm IV performs in this setting therefore require more detailed large deviations calculations that we have not carried out.

See also Collamore (2002) for a more systematic discussion of efficient estimators for unrestricted multidimensional processes.

Remark 4.5. The exponential change of measure techniques we have used exclude service times with heavy tails. In fact, it is an important problem how to perform rare events simulation efficient in this case but largely unsettled. Asmussen et al. (2000) give two logarithmically efficient algorithms for M/G/1, but also a number of counterexamples indicating the difficulty of the problem.

Basically, Algorithm IV* reduces the problem of finding a logarithmically efficient estimator for $\mathbb{P}(\tau(x) \leqslant T(x))$ in the heavy-tailed case to finding one for $\mathbb{P}(\tau(x) < C)$. This may not sound like a too ambituous goal but has not been done even for M/G/1.

5. Remaining proofs

Proof of Theorem 2.5. Recall that $C_1, C_2, \ldots$ are the lengths of the successive cycles and define $U(\cdot\,;x)$ as in Algorithm IV and $U(\cdot)$ by

$$U(A) = \sum_{n=0}^{\infty} I(C_1 + \cdots + C_n \in A),$$

(U is the renewal measure associated with the regeneration points). If $n(x)$ denotes the expected number of cycles before $T(x)$ (including the one straddling $T(x)$) where level x is exceeded, then $n(x) = \bar{f}(x)U(T(x))$ is of order $\bar{f}(x)T(x)/\mathbb{E}C$ so that

$$\limsup_{x\to\infty} \frac{\mathbb{P}(\tau(x) \leqslant T(x))}{\bar{f}(x)T(x)/\mathbb{E}C} \leqslant \limsup_{x\to\infty} \frac{n(x)}{\bar{f}(x)T(x)/\mathbb{E}C} \leqslant 1.$$

Conversely, if x is so large that $\bar{f}(\varepsilon T(x);x) \geqslant (1-\varepsilon)\bar{f}(x)$, then (4.2) yields

$$\begin{aligned}
&\mathbb{P}(\tau(x) \leqslant T(x)) \\
&\qquad \geqslant \int_0^{(1-\varepsilon)T(x)} \bar{f}(T(x)-t;x)U(\mathrm{d}t;x) \\
&\qquad \geqslant \int_0^{(1-\varepsilon)T(x)} \bar{f}(\varepsilon T(x);x)U(\mathrm{d}t;x) \\
&\qquad \geqslant (1-\varepsilon)\bar{f}(x)U((1-\varepsilon)T(x);x) \\
&\qquad \geqslant (1-\varepsilon)\bar{f}(x)[U((1-\varepsilon)T(x)) - n(x)]
\end{aligned}$$

and we get

$$\liminf_{x\to\infty} \frac{\mathbb{P}(\tau(x) \leqslant T(x))}{\bar{f}(x)T(x)/\mathbb{E}C} \geqslant (1-\varepsilon)\liminf_{x\to\infty} \frac{U((1-\varepsilon)T(x)) - n(x)}{T(x)/\mathbb{E}C} = (1-\varepsilon)^2 - 0.$$

Let $\varepsilon \downarrow 0$. □

Proof of Corollary 2.6. It is well known (e.g. Asmussen, 2000, IV.4) that under the conditions of Lemma 2.4, one has

$$\lim_{x\to\infty}\frac{\bar f(ax;x)}{\bar f(x)}=\begin{cases}0 & a<\kappa'(\gamma),\\ 1 & a>\kappa'(\gamma).\end{cases}$$

This implies the second condition in (2.1), and the rest is easy translation. □

Proof of Theorem 2.7. It is a standard large deviations estimate (e.g. Bucklew, 1990, pp. 9–10) that in the given setting

$$\frac{1}{x}\log\mathbb{P}(X(T(x))\geqslant x)\to-\kappa^*(m)/m.$$

From (1.2), it therefore follows that

$$\liminf_{x\to\infty}\frac{1}{x}\log\mathbb{P}(\tau(x)\leqslant T(x))\geqslant-\kappa^*(m)/m.$$

According to a well known large deviations estimate for $\omega(x)$ we also have

$$\frac{1}{x}\log\mathbb{P}(\omega(x)\leqslant T(x))\to-\kappa^*(m)/m$$

(see Asmussen, 2000, IV.4, X.4 and note that even if only the case $\kappa'(0)<0$, $m>\kappa'(\gamma)$ is treated there, the crucial feature of this assumption is $\kappa(\theta)>0$ which also holds when $m>\kappa'(0)>0$). The upper bound for lim sup therefore follows from the following Lemma 5.1. □

Lemma 5.1. *For some constants $c,y>0$, it holds for all $x\geqslant y$ that*

$$\mathbb{P}(\tau(x)\leqslant T)\leqslant c\lceil T\rceil\mathbb{P}(\omega(x-y)\leqslant\lceil T\rceil).$$

Proof. Assume for a moment that $\{X(t)\}$ has discrete time $t=0,1,2,\ldots$. Then for integer T,

$$\mathbb{P}(\tau(x)\leqslant T)\leqslant T\mathbb{P}(\omega(x)\leqslant T).\tag{5.1}$$

Indeed, $\tau(x)\leqslant T$ implies $X(m)-X(k)\geqslant x$ for some m,k with $0\leqslant k<m\leqslant T$, and hence

$$\begin{aligned}\mathbb{P}(\tau(x)\leqslant T)&\leqslant\sum_{k=0}^{T-1}\mathbb{P}\left(\sup_{m=k+1,\ldots,T}[X(m)-X(k)]\geqslant x\right)\\&\leqslant\sum_{k=0}^{T-1}\mathbb{P}\left(\sup_{m=k,\ldots,T+k+1}[X(m)-X(k)]\geqslant x\right)\\&=T\mathbb{P}(\omega(x)\leqslant T).\end{aligned}$$

In the general case, choose c,y with $\mathbb{P}(\inf_{0\leqslant t\leqslant1}X(t)\geqslant-y)\geqslant c^{-1}$. Then if $Q(t)\geqslant x$ for some t, say $k-1\leqslant t\leqslant k$, we have $Q(k)\geqslant x-y$ with probability at least c^{-1},

and hence

$$\mathbb{P}(\tau(x) \leqslant T) \leqslant \mathbb{P}\left(\sup_{t \leqslant \lceil T \rceil} Q(t) \geqslant x\right) \leqslant c\mathbb{P}\left(\sup_{k=1,\ldots,\lceil T \rceil} Q(k) \geqslant x - y\right)$$
$$\leqslant c\lceil T \rceil \mathbb{P}(\omega(x-y) \leqslant \lceil T \rceil)$$

using (5.1) in the last step. □

Remark 5.2. Theorem 2.7 can also be deduced from standard large deviations theory in the form of Mogulskii's theorem, see Dembo and Zeitouni (1998) and de Acosta (1994) (note that Dembo and Zeitouni (1998) impose the condition that $\kappa(\alpha) < \infty$ for all α which is unpleasant in the present queueing context since it excludes, say, the M/M/1 or M/PH/1 workload process; see de Acosta (1994) Section 5 for the version using only steepness that we use here).

The details go as follows. All functions f are assumed to be in $L^\infty[0,1]$ (equipped with the supremum norm), and we use the acronym AC for absolutely continuous. Define

$$\Lambda^*(f) = \begin{cases} \displaystyle\int_0^1 \kappa^*(f'(t))\,\mathrm{d}t & \text{if } f \text{ is AC,} \\ \infty & \text{otherwise,} \end{cases}$$

$$\underline{f}(t) = \inf_{0 \leqslant s \leqslant t} f(s), \quad \tau(f) = \inf\{t \in [0,1]\colon f(t) - \underline{f}(t) \geqslant m\}$$

($\tau(f) = \infty$ if no such t exists). Let $\mathbb{P}_T$ denote the distribution of $\{X(tT)/T\}_{0 \leqslant t \leqslant 1}$ in L^∞ and define $\Gamma = \{f\colon f(0) = 0, \tau(f) \leqslant 1\}$. Then $\mathbb{P}(\tau(x) \leqslant T) = \mathbb{P}_T(\Gamma)$. Since Γ is closed with

$$\mathrm{AC} \cap \partial\Gamma \subseteq \left\{f\colon \sup_{0 \leqslant t \leqslant 1} [f(t) - \underline{f}(t)] = m\right\},$$

it is easily seen that for $f \in \mathrm{AC} \cap \partial\Gamma$, there is a sequence $f_n \in \overset{\circ}{\Gamma}$ with $f_n \to f$, $\Lambda^*(f_n) \to \Lambda^*(f)$. Hence by Mogulskii,

$$\lim_{T \to \infty} \frac{1}{T} \log \mathbb{P}_T(\Gamma) = -\inf_{f \in \Gamma} \Lambda^*(f). \tag{5.2}$$

Let $f \in \mathrm{AC} \cap \Gamma$. Then by continuity, $f(\tau(f)) - f(\underline{\tau}(f)) = m$ for some $\underline{\tau}(f) < \tau(f)$, and we get

$$\Lambda^*(f) \geqslant \int_{\underline{\tau}(f)}^{\tau(f)} \kappa^*(f'(t))\,\mathrm{d}t \geqslant (\tau(f) - \underline{\tau}(f))\kappa^*\left(\frac{1}{\tau(f) - \underline{\tau}(f)} \int_{\underline{\tau}(f)}^{\tau(f)} f'(t)\,\mathrm{d}t\right)$$
$$= (\tau(f) - \underline{\tau}(f))\kappa^*\left(\frac{m}{\tau(f) - \underline{\tau}(f)}\right) \geqslant \kappa^*(m),$$

where we used $\kappa^* \geqslant 0$ in the first step, Jensen's inequality in the second and the convexity of κ^* on $[\kappa'(0), m]$ combined with $\kappa^*(\kappa'(0)) = 0$ and $\tau(f) - \underline{\tau}(f) \leqslant \tau(f) \leqslant 1$

in the last. Hence the inf in (5.2) is no smaller than $\kappa^*(m)$. On the other hand, the value $\kappa^*(m)$ is attained for $f_0 \in \Gamma$ defined by $f_0(t) = mt$. Hence the r.h.s. of (5.2) is $-\kappa^*(m)$, whereas the l.h.s. can be rewritten as $m \lim_{x\to\infty} \log \mathbb{P}(\tau(x) \leqslant T)/x$. From this the result follows.

Proof of Theorem 3.1(a). According to the estimate of the rare event probability given by Theorem 2.7, we must show that

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbb{E}_\theta Z_{\mathrm{I}}^2 \leqslant -2\kappa^*(m)/m.$$

Assume w.l.o.g. that $T(x) = x/m$. Now

$$\begin{aligned}
\mathbb{E}_\theta Z_{\mathrm{I}}^2 &= \mathbb{E}_\theta[\exp\{-2\theta X(\tau(x)) + 2\tau(x)\kappa(\theta)\}; \tau(x) \leqslant T] \\
&= \mathbb{E}_\theta[\exp\{-2\theta Q(\tau(x)) + 2\theta L(\tau(x)) + 2\tau(x)\kappa(\theta)\}; \tau(x) \leqslant T] \\
&\leqslant \mathbb{E}_\theta[\exp\{-2\theta x + 2\theta L(\tau(x)) + x\kappa(\theta)/m\}; \tau(x) \leqslant T] \\
&= \mathrm{e}^{-2x\kappa^*(m)/m} \mathbb{E}_\theta[\mathrm{e}^{2\theta L(\tau(x))}; \tau(x) \leqslant T]. \qquad (5.3)
\end{aligned}$$

To complete the proof, it therefore suffices to show that $\mathbb{E}_\theta \mathrm{e}^{2\theta L(\infty)}$ is finite when $\kappa(-\theta) < \kappa(\theta)$. Now it is well known that for any Lévy process with $\kappa'(0) < 0$ and κ steep, we have

$$\mathbb{E} \exp\left\{\alpha \sup_{0\leqslant t<\infty} X(t)\right\} < \infty \quad \text{for all } \alpha < \gamma.$$

Changing the sign shows that if instead $\kappa'(0) > 0$, then $\mathbb{E}\,\mathrm{e}^{\alpha L(\infty)}$ is finite for all $\alpha < \eta$ where η satisfies $\kappa(-\eta) = 0$. Translating to the process with Lévy exponent κ_θ, this means that $\mathbb{E}_\theta \mathrm{e}^{\alpha L(\infty)}$ is finite for all $\alpha < \eta$ where η satisfies $\kappa(\theta - \eta) = \kappa(\theta)$. But since $\kappa(-\theta) < \kappa(\theta)$, we have $-\theta > \theta - \eta$, i.e. $\eta > 2\theta$ so that $\mathbb{E}_\theta \mathrm{e}^{\alpha L(\infty)}$ is finite when $\alpha = 2\theta$. □

Proof of Theorem 3.1(b). We first present a heuristical argument. The idea is to first note that reversing the estimates around (5.3) shows that

$$\mathbb{E}_\theta Z_{\mathrm{I}}^2 \approx \mathrm{e}^{-2x\kappa^*(m)/m} \mathbb{E}_\theta[\mathrm{e}^{2\theta L(\tau(x))}; \tau(x) \leqslant T]$$

and next to look for a path $\{x_0(s)\}_{0\leqslant s\leqslant 1}$ of $\{X(sT)/T\}_{0\leqslant s\leqslant 1}$ which makes $L(\tau(x))$ large, say of order Tty for some $y > 0$ and some $t \in (0,1)$, at the same time as it makes $\mathbb{P}_\theta(\{X(\cdot T)/T\} \approx x_0)$ no smaller than that $\mathrm{e}^{2\theta y}\mathbb{P}_\theta(\{X(\cdot T)/T\} \approx x_0, \tau(x) \leqslant T)$ is large. We will see that this is achieved by requiring the drift to be changed from m to $-y$ in the time interval $(0, Tt]$ and to $m/(1-t)$ in the time interval $(Tt, T]$, cf. Fig. 2 (t and y will be specified later). That is, x_0 is the function whose graph connects the points $(0,0)$, $(t, -ty)$, $(1, m - ty)$ linearly.

Now it is well known that the overshoot $\xi_Q(x)$ is bounded in distribution and hence by (5.3), it should hold that

$$\mathbb{E}_\theta Z_{\mathrm{I}}^2 \approx c\mathbb{E}_\theta[\exp\{-2\theta x + 2\theta L(\tau(x)) + 2\tau(x)\kappa(\theta))\}; \tau(x) \leqslant T].$$

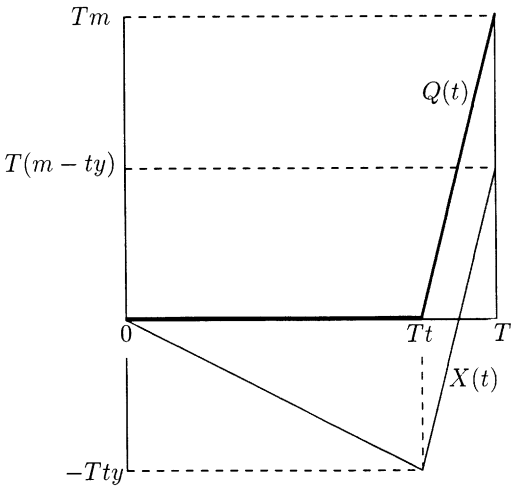

Fig. 2.

Using Mogulskii in the form

$$\log \mathbb{P}_\theta(X(sT)/T \approx x_0(s),\ s \leqslant 1) \sim -tT\kappa^*(-y) - (1-t)T\kappa_\theta^*\left(\frac{m}{1-t}\right)$$

and replacing $\{Q(T) \geqslant x\}$ by $\{\{X(sT)/T\} \approx x_0(s),\ s \leqslant 1\}$ and $\tau(x)$ by T yields

$$\liminf_{T\to\infty} \frac{1}{T} \log \mathbb{E}_\theta Z_1^2 \geqslant -2\kappa^*(m) + 2\theta ty - t\kappa_\theta^*(-y) - (1-t)\kappa_\theta^*\left(\frac{m}{1-t}\right). \quad (5.4)$$

For a complete proof, we therefore need to show that

$$2\theta ty - t\kappa_\theta^*(-y) - (1-t)\kappa_\theta^*\left(\frac{m}{1-t}\right) > 0 \quad (5.5)$$

for an appropriate choice of y and t when $\kappa(-\theta) > \kappa(\theta)$, and to verify (5.4) rigorously.

Write $\alpha(x;\kappa)$ for the solution of $\kappa'(\alpha) = x$ and so on. From $\alpha(x;\kappa_\theta) = \alpha(x;\kappa) - \theta$, we get

$$\begin{aligned}\kappa_\theta^*(x) &= \alpha(x;\kappa_\theta)x - \kappa_\theta(\alpha(x;\kappa_\theta)) = (\alpha(x;\kappa) - \theta)x - \kappa(\alpha(x;\kappa)) + \kappa(\theta)\\ &= \kappa^*(x) - \theta x + \kappa(\theta) = \kappa^*(x) - \kappa^*(m) + \theta(m-x)\end{aligned}$$

from which we conclude that $\alpha(\beta;\kappa_\theta^*) = \alpha(\beta+\theta;\kappa^*)$. Write $\tilde{\kappa}(\theta) = \kappa(-\theta)$ and so on. Then

$$\alpha(x;\tilde{\kappa}) = -\alpha(-x;\kappa),\ \tilde{\kappa}^*(x) = \kappa^*(-x).$$

It follows that the expression $2\theta y - \kappa_\theta^*(-y) = 2\theta y - \tilde{\kappa}_\theta^*(y)$ is maximized by taking

$$y = \alpha(2\theta; \tilde{\kappa}_\theta^*) = -\alpha(-2\theta; \kappa_\theta^*) = -\alpha(-\theta; \kappa^*)$$

and that the maximum value is

$$\tilde{\kappa}_\theta^{**}(2\theta) = \tilde{\kappa}_\theta(2\theta) = \kappa_\theta(-2\theta) = \kappa(-\theta) - \kappa(\theta) > 0.$$

Further, it is standard that $\kappa_\theta'(0) = m$ implies that κ_θ^* attains its minimum 0 at m so that $\kappa_\theta^*(m/(1-t)) \sim (m^2/2)t^2\kappa_\theta^{*\prime\prime}(m)$ as $t \downarrow 0$. It follows that indeed (5.5) holds for some small but positive t.

The proof of (5.4) is similar to the verification of the lower bound in Varadhan's integral lemma (Dembo and Zeitouni, 1998; Glasserman et al., 1999) (the result cannot be applied directly because of difficulties with continuity of say $\tau(f)$ on L^∞). Let $\underline{x}_\varepsilon$, $\bar{x}_\varepsilon$ be the functions whose graph connects the points $(0,0)$, $(t, -ty-\varepsilon)$, $(1, m-ty+\varepsilon)$, resp. $(0,0)$, $(t, -ty+\varepsilon)$, $(1, m-ty+2\varepsilon)$, linearly and let Γ_ε be the open set

$$\{f\colon \underline{x}_\varepsilon(s) < f(s) < \bar{x}_\varepsilon(s) \text{ for all } s \in [0,1]\}.$$

It is readily checked that for $f \in \Gamma_\varepsilon$,

$$1 - \delta(\varepsilon) \leqslant \tau(f) \leqslant 1 \quad \text{where } \delta(\varepsilon) = (1-t)\left(1 - \frac{m-2\varepsilon}{m+\varepsilon}\right).$$

Hence

$$\begin{aligned}
\mathbb{E}_\theta Z_{\mathrm{I}}^2 &= \mathbb{E}_\theta[\exp\{-2\theta X(\tau(x)) + 2\tau(x)\kappa(\theta)\}; \tau(x) \leqslant T] \\
&\geqslant \mathbb{E}_\theta[\exp\{-2\theta X(\tau(x)) + 2\tau(x)\kappa(\theta)\}; \{X(\cdot T)/T\} \in \Gamma_\varepsilon] \\
&\geqslant \exp\{-2T\theta(m - ty + 2\varepsilon) + 2T\kappa(\theta)[1-\delta(\varepsilon)]\}\mathbb{P}_\theta(\{X(\cdot T)/T\} \in \Gamma_\varepsilon)
\end{aligned}$$

and Mogulskii yields

$$\liminf_{T\to\infty} \frac{1}{T} \log \mathbb{E}_\theta Z_{\mathrm{I}}^2 \geqslant -2\kappa^*(m) + 2\theta t y - 4\theta\varepsilon - 2\kappa(\theta)\delta(\varepsilon) - \inf_{f\in\Gamma_\varepsilon} \Lambda^*(f).$$

Since it is straightforward to check that

$$\limsup_{\varepsilon\downarrow 0} \inf_{f\in\Gamma_\varepsilon} \Lambda^*(f) \leqslant \Lambda^*(x_0),$$

it follows that

$$\liminf_{T\to\infty} \frac{1}{T} \log \mathbb{E}_\theta Z_{\mathrm{I}}^2 \geqslant -2\kappa^*(m) + 2\theta t y - \Lambda^*(x_0),$$

which is the same as (5.4). □

Proof of Corollary 3.2. Since $\kappa'(0) > 0$, we have $\kappa(\theta) > 0$ for all $\theta > 0$ and $\kappa(-\theta) < 0$ for all small $\theta > 0$. Hence $\kappa(-\theta) - \kappa(\theta) < 0$ for all small $\theta > 0$. □

Proof of Corollary 3.3. Since $\gamma > 0$ and $\kappa(-\theta) > 0$ for all $\theta > 0$ because $\kappa'(0) < 0$, the function $\kappa(-\theta) - \kappa(\theta)$ is strictly positive for $\theta = \gamma$ and hence in an interval of the form (γ, γ_2) with $\gamma_2 > \gamma$.

For Brownian motion with drift $-\mu < 0$ and $\sigma^2 = 1$, we have $-\mu + \theta = m$ and

$$\kappa(-\theta) - \kappa(\theta) = \left[\frac{\theta^2}{2} + \mu\theta\right] - \left[\frac{\theta^2}{2} - \mu\theta\right] = 2\mu\theta > 0$$

for all $\theta > 0$, in particular when $m > \kappa'(\gamma)$. For the M/M/1 queue length with $\lambda < \mu$, write $z = e^{\theta}$. Since $\kappa'(\theta) = m > \kappa'(\gamma) > 0$ implies $\theta > \gamma > 0$, we have $z > 1$ and hence

$$\kappa(-\theta) - \kappa(\theta) = \left[\lambda\left(\frac{1}{z} - 1\right) + \mu(z-1)\right] - \left[\lambda(z-1) + \mu\left(\frac{1}{z} - 1\right)\right]$$

$$= (\lambda - \mu)\left(\frac{1}{z} - z\right),$$

which is strictly positive since each factor is strictly negative. □

The function $\kappa(-\theta) - \kappa(\theta)$ is the difference between two convex functions and hence it seems difficult to say something about the set of θ's where it is negative (Algorithm I is logarithmic efficient). The intricacies may become clear if one compares Corollary 3.3 with the following example:

Example 5.3. Consider the M/G/1 workload process (Example 2.2), and assume that $\mathbb{E}\, e^{\theta U}$ has the finite radius θ_+ of convergence (the steepness assumption then means $\mathbb{E}\, e^{\theta U} \uparrow \infty$ as $\theta \uparrow \theta_+$; for example, U could be exponential or Gamma). In the stable case $\kappa'(0) < 0$, $\kappa(-\theta) \leqslant \kappa(-\theta_+)$ for all $\theta \leqslant \theta_+$, and it follows that there exists $\theta_3 \in (\gamma, \theta_+)$ such that $\kappa(-\theta) - \kappa(\theta) < 0$ when $\theta \in (\theta_3, \theta_+)$. In fact, as upper bound for θ_3 one may take the solution θ_4 of $\kappa(\theta_4) = \kappa(-\theta_+)$, and logarithmic efficiency holds then at least when $m > \kappa'(\theta_4)$.

6. Concluding remarks

1. All of our algorithms using exponential change of measure applies with small changes to Markov-modulated queues where $\{Q(t)\}$ is the reflected version of an additive process $\{X(t)\}$ on a finite Markov process $\{J(t)\}$. In this case, the exponential change of measure with parameter θ is performed by first determining the matrix $K[\theta]$ such that the matrix with ijth entry

   $$\mathbb{E}[e^{\theta X(t)}; J(t) = j \,|\, J(0) = i]$$

   has the form $e^{tK[\theta]}$. One then defines $\kappa(\theta)$ as the eigenvalue with largest real part of $K[\theta]$. Letting h be the corresponding right eigenvector, the likelihood ratio up to T is

   $$\left.\frac{d\mathbb{P}}{d\mathbb{P}_\theta}\right|_T = \exp\{-\theta X(T) + T\kappa(\theta)\}\,\frac{h_{J(0)}}{h_{J(T)}}.$$

   See Asmussen and Rubinstein (1995) for further details.
2. As an example related to but simpler than the problem of this paper, it is instructive to consider the case $A(x) = \{Q(T) \geqslant x\}$ of a large queue length at time T.

The conditional distribution of $\{Q(t)\}_{0\leqslant t\leqslant T}$ given $A(x)$ is described in Anantharam (1988). The result is that

(a) If $T\kappa'(\gamma) < x$, then the $\mathbb{P}(\cdot\,|\,A(x))$–distribution of $\{Q(t)\}_{0\leqslant t\leqslant T}$ is close to the $\mathbb{P}_\theta$-distribution where θ is chosen according to a linear drift from level 0 at time $t=0$ to level x at time T, i.e., θ is determined by the equation $T\kappa'(\theta)=x$ (note that this implies $\theta>\gamma$). That is, for the simulation one would as a first attempt use the corresponding exponential change of measure on the whole of $[0,T]$.

(b) If $T\kappa'(\gamma) > x$, then the $\mathbb{P}(\cdot\,|\,A(x))$-distribution of $\{Q(t)\}_{0\leqslant t\leqslant T}$ is instead close to the $\tilde{\mathbb{P}}$–distribution where $\tilde{\mathbb{P}}$ is the distribution such that $\{Q(t)\}_{0\leqslant t\leqslant t(x)}$ develops normally with parameters λ,μ and $\{Q(t)\}_{t(x)\leqslant t\leqslant T}$ with parameters $\lambda_\gamma=\mu,\ \mu_\gamma=\lambda$ where $t(x)=T-x/\kappa'(\gamma)$. Thus in the simulation one would as a first attempt start 'normally' and first switch on the change of measure (corresponding to $\mathbb{P}_\gamma$) at time $t(x)$.
The results of this paper indicate, however, that these ideas do not work and that one needs to modify along the lines of Algorithms IV, IV*.

7. Uncited reference

Glasserman and Kou, 1995b

References

Anantharam, V., 1988. How large delays build up in a *GI*/*GI*/1 queue. Queueing Syst. 5, 345–368.

Asmussen, S., 1982. Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the *GI*/*G*/1 queue. Adv. Appl. Probab. 14, 143–170.

Asmussen, S., 1987. Applied Probability and Queues, 2nd Edition. Wiley, Chichester (to be published Springer, New York, 2002).

Asmussen, S., 1998. Subexponential asymptotics for stochastic processes: extremal behavior, stationary distributions and first passage probabilities Ann. Appl. Probab. 8, 354–374.

Asmussen, S., 1999. Extreme value theory for queues via cycle maxima. Extremes 1, 137–168.

Asmussen, S., 2000. Ruin Probabilities. World Scientific, Singapore.

Asmussen, S., Kella, O., 2001. On optional stopping of some exponential martingales for Lévy processes with or without reflection. Stoch. Proc. Appl. 91, 47–55.

Asmussen, S., Rubinstein, R.Y., 1995. Steady-state rare events simulation in queueing models and its complexity properties. In: Dshalalow, J. (Ed.), Advances in Queueing: Models, Methods & Problems. CRC Press, Boca Raton, FL, pp. 429–466.

Asmussen, S., Binswanger, K., Højgaard, B., 2000. Rare events simulation for heavy-tailed distributions. Bernoulli 6, 303–322.

Asmussen, S., Jobmann, M., Schwefel, H.-P., 2002. Exact buffer overflow calculations for queues via martingales. Queueing Syst., in preparation.

Bertoin, J., 1990. Lévy Processes. Cambridge University Press, Cambridge.

Bratley, P., Fox, B.L., Schrage, L., 1987. A Guide to Simulation. Springer, New York.

Bucklew, J.A., 1990. Large Deviation Techniques in Decision, Simulation and Estimation. Wiley, New York.

Bucklew, J.A., Ney, P., Sadowsky, J.S., 1990. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. J. Appl. Probab. 27, 44–59.

Collamore, J.F., 2002. Importance sampling techniques for the multidimensional ruin problem for general Markov additive sequences of random vectors. Ann. Appl. Probab. 12, 382–421.

de Acosta, A., 1994. Large deviations for vector-valued Lévy processes. Stoch. Proc. Appl. 51, 75–115.

Dembo, A., Zeitouni, O., 1998. Large Deviations Techniques and Applications, 2nd Edition. Springer, Berlin.

Frantz, P., 2000. Efficient Techniques for Simulating Telecommunication Systems. Masters Thesis, Department of Computer Science, Technical University of Munich.

Glasserman, P., 1996. Filtered Monte Carlo. Math. Oper. Res. 18, 610–634.

Glasserman, P., Kou, S.-G., 1995a. Analysis of an importance sampling estimator for tandem queues. ACM TOMACS 4, 22–42.

Glasserman, P., Kou, S.-G., 1995b. Limits of first passage times to rare sets in regenerative processes. Ann. Appl. Probab. 5, 424–445.

Glasserman, P., Wang, Y., 1997. Counterexamples in importance sampling for large deviations probabilities. Ann. Appl. Probab. 7, 731–746.

Glasserman, P., Heidelberger, P., Shahabuddin, P., 1999. Asymptotically optimal importance sampling and stratification for pricing path-dependent options. Math. Finance 9, 117–152.

Gnedenko, B.V., Kovalenko, I.N., 1989. Introduction to Queueing Theory, 2nd Edition. Birkhäuser, Basel.

Heidelberger, P., 1995. Fast simulation of rare events in queueing and reliability models. ACM TOMACS 6, 43–85.

Keilson, J., 1966. A limit theorem for passage times in ergodic regenerative processes. Ann. Math. Statist. 37, 866–870.

Kella, O., Whitt, W., 1992. Useful martingales for stochastic storage processes with Lévy input. J. Appl. Probab. 29, 396–403.

Whitt, W., 2002. Stochastic Process Limits. Springer, New York.